DOCUMENT RESUME

ED 225 049 CG 016 397

AUTHOR Wolf, Fredric M.
TITLE Meta-Analytic Applications in Program Evaluation.
PUB DATE Aug 82
NOTE 27p.; Paper presented at the Annual Covention of the American Psychological Association (90th, Washington, DC, August 23-27, 1982).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Case Studies; *Data Analysis; Elementary Secondary Education; *Evaluation Methods; Higher Education; Literature Reviews; Pretests Posttests; *Program Effectiveness; *Program Evaluation; Psychological Evaluation; *Research Methodology; Student Evaluation
IDENTIFIERS *Meta Analysis

ABSTRACT

In a variety of psychological and educational situations, it is desirable to be able to make data-based evaluative summary statements regarding the impact of a given program. Certain procedures typically used in meta-analytic studies that review and integrate results from individual studies, such as combined tests and measures of effect size, are particularly well suited for program evaluation in certain situations. This paper describes a number of such situations, briefly reviews the literature on combined tests and effect size, and provides several illustrative numerical examples of their application in program evaluation. The three examples illustrate the practical utility of using combined tests and measures of effect size in program evaluations in situations where data are available either cross-sectionally, or on successive occasions, or on independent components of a larger program. The materials suggest that measures of effect size are clearly valuable in providing potential insight into the differential impact of a given program, information that is more obscured when relying solely on statistical tests. (Author/JAC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED225049

Presented at the meeting of The
American Psychological Association,
Washington, D.C., August 1982.

Meta-Analytic Applications in Program Evaluation

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Fredric M. Wolf

University of Michigan

Medical School

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Fredric M. Wolf

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Running head: Meta-Analytic Evaluation

CG 016397

Address requests for reprints to Fredric M. Wolf, Department of Postgraduate Medicine and Health Professions Education, University of Michigan Medical School, G1208 Towsley Center, Ann Arbor, Michigan 48109.

Meta-Analytic Applications in Program Evaluation

Arriving at data-based summary statements regarding the effectiveness of a given program is of considerable potential value for interpreting the outcomes of psychological and educational evaluations. For example, an evaluator may wish to integrate the independent outcome results of a program implemented cross-sectionally across various age or grade levels. In another situation, an evaluator may wish to integrate the results of a program implemented with independent samples of similar subjects over successive time periods, such as quarters, semesters, or years. In still another situation, an evaluator may wish to integrate the results of various independent services that an educational or social service agency provides. These three situations will be referred to as the a) cross-sectional, b) independent samples/similar subjects - successive occasions, and c) independent program components cases, respectively. Certain procedures, such as combined tests and measures of effect size, that are typically used in meta-analytic studies to statistically integrate the findings of a large collection of results from individual studies, are particularly well suited for program evaluation in these situations.

The purpose of the present paper is to briefly review the recent literature on combined tests and effect size, indicate how they may be used effectively in program evaluation, and provide several illustrative numerical examples of their actual application in program evaluation. It should be understood that these procedures are distinct from those known as meta evaluation (Cook & Gruder, 1979), which denotes the evaluation of evaluations.

Combining Results of Independent Tests

Statistical methods available for combining the results of independent studies range from various counting procedures to a variety of summation procedures involving either significance levels (probabilities or their logarithmic transformations) or raw or weighted test statistics such as ts or zs.

Since R.A. Fisher (1932) and Karl Pearson (1933) independently addressed the issue of statistically summarizing the results of independent tests of the same hypothesis, interest in these types of procedures has continued. More recently this process has been called meta-analysis (Glass, 1976) for "statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p. 3). For a thorough review of the "traditional" meta-analysis approach to the review and synthesis of research literature, the reader is referred to Glass (1976, 1978) and Glass, McGaw, and Smith (1981). The present paper addresses the application of these procedures to program evaluation rather than to the synthesis of research literature on a given topic.

These procedures have become known as "combined tests," and have been illustrated by Rosenthal (1978) and Winer (1971), among others. While a variety of tests for combining the results of independent tests of the same hypothesis have been put forward (see Birnbaum, 1954; Rosenthal, 1978; Van Zwet and Oosterhoff, 1967 for reviews of these tests), only the procedures presented by Fisher (1932, 1948), Winer (1971), and Stouffer (1949; Mosteller & Bush, 1954) will be discussed in the present paper.

In addressing the question of combining the results of a number of independent tests which have all been planned to test a common hypothesis, Fisher described a method based on the product of probabilities from different trials. If the natural logarithms of these probabilities are calculated and then multiplied by minus two (-2) and then summed, a chi square with degrees of freedom equal to two times the number of tests combined (2n) is obtained (the logarithmic transformation permits a summative rather than a multiplicative function, thereby simplifying calculations). This may be expressed in the form of

$$\chi^2 = -2 \Sigma \log_e p, \qquad (1)$$

with df = 2n

where n = number of tests combined

and p = one tailed probability associated with each test.

This procedure has been shown to be more efficient than several of the other combining methods (Koziol & Perlman 1978; Littell & Folks 1973), although it suffers from several limitations (Rosenthal 1978). Mosteller and Bush (1954) noted that it can yield results inconsistent with a simple sign test in situations where the majority of a large number of studies showed results in one direction with p values close to .50 (i.e. chance). In this situation the sign test could easily reject the overall null hypothesis, while the Fisher procedure would not. The Fisher procedure would thus yield more conservative results in this situation, a result not terribly disturbing given the recent recommendations of reporting the effect size as well as the overall probability level when using combined tests (McGaw & Glass 1980; Rosenthal 1978). That is, while the sign test would be significant in this instance, the effect size would likely be small and thus more appropriately tested with the Fisher method which would result in non-significance.

A more serious disadvantage of the Fisher procedure, however, is its support for the significance of either outcome when two studies of equally and strongly significant results in opposite directions are obtained (Adcock, 1960). Even given these limitations, this procedure remains one of the best known and applied.

Winer (1971) has presented a procedure for combining independent tests that comes directly from the sampling distribution of independent t-statistics in which the t-statistics associated with each test are summed and divided by the square root of the sum of the degrees of freedom (df) associated with each t after each df has been divided by df -2. This may be expressed in the form of

$$z = \frac{\Sigma\ t}{\sqrt{\Sigma\ [df/(df - 2)]}} \qquad (2)$$

This procedure is based on df/(df - 2) being the variance of a t distribution, which is approximately normally distributed (N (O,1)) when df $\geq$ 10. Thus this procedure is not appropriate for tests based on very small samples (less than 10) and, as Rosenthal (1978) pointed out, "cannot be employed at all when the size of the samples for which t is computed becomes less than three, because that would involve dividing by zero or by a negative value." In practice, however, it is not common for tests of significance to be applied to such small samples, thereby minimizing the effect of this disadvantage.

A third approach originally attributed to Stouffer (1949) is more fully described by Mosteller and Bush (1954) and Rosenthal (1978). It is similar to the Winer procedure of summing t's, with the exception that p values are converted to zs instead of to ts, and then summed. The denominator then simplifies to the square root of the number of tests combined, and the complete expression takes the form of

$$z = \frac{\Sigma\ z}{\sqrt{N}} \qquad (3)$$

where N= number of tests combined. This procedure is based on the sum of normal deviates being itself a normal deviate, with the variance equal to the number of observations summed.

The Stouffer procedure offers several advantages. The calculations are more straightforward than both the Fisher procedure, which necessitates logarithmic transformations, and the Winer procedure, which makes an adjustment for degrees of freedom. In addition, results of the z procedure, while slightly more powerful, are virtually identical to results of the t procedure (Wolf & Spies, 1981). This is particularly true when the statistics summed are derived from large samples, as df/df-2 approaches unity as sample size increases.

Measuring Effect Size

Glass' exposition and application of meta-analysis relies heavily on the use of measures of effect size that have been eloquently summarized by Cohen (1977). Cohen states, "Without intending any necessary implication of causality, it is convenient to use the phrase 'effect size' to mean 'the degree to which the phenomenon is present in the population', or 'the degree to which the null hypothesis is false'. Whatever the manner of representation of a phenomenon in a particular research in the present treatment, the null hypothesis always means that the effect size is zero" (pp. 9-10).

Statistical tests such as the combined procedures previously described provide a summary index of the statistical significance of the results pertaining to an hypothesis. They do not, however, provide any insight into the strength of the relationship or effect of interest. The desirability of accompanying combined tests with indexes of effect size has been noted by Rosenthal (1978). McGaw and Glass (1980) and Glass, McGaw, and Smith (1981) provide helpful guidelines for converting various summary statistics into a common metric,

usually in the form of the Pearson Product Moment Correlation. Cohen (1977) provides measures of effect size for most common statistical tests. Because many program evaluations consist of pre-post and/or experimental-control group designs, the effect size measures for t-tests between means will be illustrated here. The reader is referred to the above references for measures of effect size appropriate for other statistical tests.

The goal is to obtain "a pure number, one free of our original measurement unit, with which to index what can be alternatively called the degree of departure from the null hypothesis of the alternative hypothesis, or the ES (effect size) we wish to detect. This is accomplished by standardizing the raw effect size as expressed in the measurement unit of the dependent variable by dividing it by the (common) standard deviation of the measures in their respective populations, the latter also in the original measurement" (Cohen, 1977, p. 20).

This may be accomplished in the form of

$$d = \frac{|\bar{x}_1 - \bar{x}_2|}{\sigma} \qquad (4)$$

where d = ES index for t-tests of means in standard unit, $\bar{x}_1$ and $\bar{x}_2$ = sample means in original measurement units, and σ =standard deviation of either sample (as homogeneity of variance is assumed).

The means, $\bar{x}_1$ and $\bar{x}_2$, are typically the experimental and control group means in posttest-only control group experimental designs, or pre and post means in one group pretest-posttest pre-experimental designs. It should be noted that the latter design may be considered "primitive yet adequate if the treated group members' pretreatment status is a good estimate of their hypothetical post-treatment status in the absence of treatment" (Glass, 1978). This is an empirical question that can be studied to determine if maturation,

pre-test sensitization or other threats to the validity of this design have in fact biased this estimate. In fact, Campbell (1982) has recently indicated that the one group, pretest-posttest design has "now been elevated to a useful quasi-experimental or proto-experimental design" in the planned revision of his classic work on research design (Campbell & Stanley, 1963).

The standard deviation, σ, is typically either the control group or pretest standard deviation, as it is assumed that the two group variances are equal. Another possibility would be to use the within population standard deviation.

Once the effect size, d, is determined, Cohen provides tables to translate d into measures of nonoverlap (U) between the two groups, which translate rather nicely into graphical displays which facilitate interpretation of the results. Perhaps the most useful index of nonoverlap is Cohen's U_3, which translates average performance in percentiles (area under the normal curve) of the experimental (or posttest) group to the equivalent percentile of the control (or pretest) group. This will be illuminated with the following numerical illustrations.

Some Illustrative Examples

The following numerical examples provide concrete illustration of these computational methods. To consolidate the various examples, all three illustrations use one group pretest-posttest designs, as these were the designs of the actual programs evaluated. Obviously, the computations would be the same if a posttest-only control group experimental design had been used, with the control group mean replacing the pretest mean and the experimental group mean replacing the posttest mean. In this instance the control group standard deviation would be used instead of the pretest standard deviation.

A decision rule that will be employed throughout is to use no significance level less than .001, two-tailed or .0005, one-tailed. This convention leads to a more conservative result when p values rather than the raw test statistics are used. In addition, it should be noted that one-tailed tests are used with combined tests (Fisher, 1932; Rosenthal, 1978, 1980; Winer, 1971), inasmuch as the results of the prior independent studies are known and the direction of the hypothesis should therefore be clear.

Case A: Cross-Sectional Synthesis

Alternate forms of the Comprehensive Tests of Basic Skills (CTB/McGraw Hill, 1973/1975) were administered under standard conditions in October at the beginning of the school year and again in May at the close of the year to 2,630 students in Grades 1 to 8 from all three elementary and both middle schools in a rural midwestern community of approximately 20,000 inhabitants. The CTBS mathematics subscales were used as part of the evaluation of a federally funded mathematics program (Wolf & Blixt, 1979, 1981). A one group pretest-posttest pre-experimental design was used to assess the change in mathematics achievement at each grade level. Results of paired t-tests summarized in Table 1 indicated that students at each grade level exhibited significant ($p < .001$, two-tailed tests) improvement at each grade level (paired $t = 14.17$ to 43.42).

- - - - - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - - - - -

Combining the results of all 8 of these independent tests of the research hypothesis (Table 2) that students would exhibit significant gains in their mathematics achievement in order to make one summary statement by applying the Fisher procedure described in formula 1 would result in:

$$\chi^2 = 15.2 + 15.2 + 15.2 + 15.2 + 15.2 + 15.2 + 15.2 + 15.2 = 121.6 \qquad (5)$$

Because there are eight independent tests of this hypothesis, one for each grade level, there are 2n or (2) (8) = 16 degrees of freedom. The critical value for an alpha level of .001 with 16 df, one-tailed is 39.25. Not surprisingly, the combined evidence from the eight tests indicates that the research hypothesis of significant gains in achievement is supported when the scope of the inference is with respect to the combined populations.

- - - - - - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - - - - - -

Similarly, when applying formula 2 for the Winer procedure to the same data, the following result is obtained:

$$z = \frac{43.32 + 35.47 + 36.11 + 24.39 + 19.24 + 17.30 + 18.04 + 14.17}{\sqrt{\frac{308}{306} + \frac{361}{359} + \frac{362}{360} + \frac{339}{337} + \frac{330}{328} + \frac{303}{301} + \frac{321}{319} + \frac{298}{296}}} \qquad (6)$$

$$= \frac{208.04}{2.84} = 73.25$$

The probability of obtaining this value of z or one larger is p ($z \geq 73.25$) $< .001$, one-tailed.

Analogous results are also obtained when formula 3 for the Stoufier procedure is applied to the data. In this approach, however, the one-tailed p values are converted to their analogous z - statistics and then summed and divided by the square root of the number of tests summed:

$$z = \frac{3.3 + 3.3 + 3.3 + 3.3 + 3.3 + 3.3 + 3.3 + 3.3}{\sqrt{8}} = \frac{26.4}{2.83} = 9.33 \qquad (7)$$

The probability of obtaining this value of z or larger is p ($z \geq 9.35$) $< .001$, one-tailed. Because a decision rule not to use p values less than .0005, one-tailed was implimented, it is noteworthy that when these p values are converted

to z-statistics, the resultant z-statistics are markedly lower than the t-statistics obtained from the original raw data. However, the overall result is equivalent.

Given that the differences between the pre and posttest means were highly significant at each grade level, it is hardly surprising that the overall combined test is also highly significant. In this instance, the magnitude of the effect may be of more practical importance and interest. Applying the effect size formula for d in equation 4 to the data for students in the first grade provides the following result:

$$d = \frac{|0.9 - 2.4|}{0.62} = \frac{1.5}{0.62} = 2.42 \qquad (8)$$

Cohen (1977) provides interpretative guidelines for effect size, with d = .2 indicative of a small effect, d = .5 indicative of a medium effect, and d = .8 indicative of a large effect. Clearly the effect for first graders in the example is a large one. Another way of conveying the same conclusion is to translate d into a measure of overlap (U). Cohen (1977) provides tables for making this transition, although values obtained from a normal distribution table are essentially equivalent to Cohen's U_3 tabled values. A d value of 2.42 translates into a U_3 value of .992. This means that the average score (50th percentile) on the posttest was equivalent to the 99.2nd percentile on the pretest.

The effect size would be calculated in a similar fashion for each of the other seven grade levels. These individual effect sizes typically are averaged to obtain the mean effect size over all grade levels, which in the present instance was 1.19. This average effect size translates into a U_3 value of .838. Thus across all eight grade levels we could expect the average performance on the posttest to be equivalent to the 83.8th percentile on the pretest. This is presented graphically in figure 1.

Insert Figure 1 about here

Interestingly, however, effect sizes ranged from a high of 2.42 (large effect) for first graders to a low of 0.52 (medium effect) for eighth graders, with a generally downward trend with increasing age (grade level). An examination of means and standard deviations for individual grade levels suggests that this decreasing trend is a result, in part, of the increasing variance associated with increasing grade levels. This could perhaps suggest that individual differences in mathematics achievement are relatively homogeneous at the beginning of formal education, but become much more pronounced with greater educational experience. This in turn suggests that the program on the average had greatest impact in the earlier grades, even though the impact was quite noticeable throughout.

It is noted, however, that these interpretations are very speculative given the nature of pre-post designs. That is, threats to the validity of these results through maturation and normal academic progress effects (not resulting from this specific treatment program) are uncontrolled in this design. A more appropriate evaluation design would be to compare the performance of each grade after it had the program with that of the same grade for the previous year, which didn't. This would then confound the program treatment effects with only historic and cohort differences. The same combined test and effect size procedures could then be performed on this non-equivalent control group design as were presented here. The present example was presented only for illustrative purposes.

Case B: Independent Samples/Similar Subjects-Successive Occasions Synthesis

First-year medical students participated in a 10 week course designed to improve their communication and interviewing skills (Engler, Saltzman, Walker & Wolf, 1981; Saltzman, Wolf, Savickas & Walker, 1981; Wolf, 1981). A Standard Index of Communication (Carkhuff, 1969) was administered both before and after training in a one sample pretest-posttest design. The first three successive classes of students each exhibited significant gains on this Index, which rates students' responses to a series of patient situations/statements. While it is important to monitor each class' performance independently, summarizing the results across all samples of similar subjects who participated in the program during successive academic years provides a more stable estimate of the effectiveness of training.

All three classes exhibited significant gains (paired $t = -8.55$ to -24.18; $p < .001$, two-tailed). The Fisher ($\chi^2(6) = 45.6$), Winer ($z = 26.43$), and Stouffer ($z = 5.72$) combined tests were each highly significant ($p < .001$, one-tailed). The average effect size was 2.90 σ_x (Sd = 0.75 σ_x), indicating that the average performance at posttesting was equivalent to the 99.8th percentile on the pretest. These findings are summarized in tables 3 and 4.

Insert Tables 3 and 4 about here

Case C: Synthesizing Independent Program Components

Goal attainment scaling (Kiresuk & Lund, 1976) was used to evaluate the impact of services provided by four independent agencies (Adult Mental Health

Center, Elderly Home Aid Services, Crisis Intervention/Hotline, and Children's Services) that comprise a county mental health board (Wolf & Blixt, 1981). Goal attainment follow-up guides were completed at intake and again during follow-up 10 weeks later. Paired t-tests indicated that on the average clients in each of the agencies exhibited significant improvement (paired $t = 10.28$ to 12.02; $p < .001$, two-tailed). Combined tests used to synthesize and summarize these independent results confirmed ($p < .001$, one-tailed) these findings with respect to the combined populations (Fisher $\chi^2 = 60.8$; Winer $z = 20.91$; Stouffer $z = 6.60$). The average effect size of 3.79 σ_x indicated that average follow-up scores were equivalent to scores at the 99.9th percentile on the distribution of scores at intake. These findings are summaried in more detail in tables 5 and 6.

Insert Tables 5 and 6 here

Conclusions and Recommendations

The above examples illustrate the practical utility of using combined tests and measures of effect size in program evaluation in situations where data are available either cross-sectionally, or on successive occasions, or on independent components of a larger program. It is suggested that a combined test and measure of effect size both be used rather than presenting one without the other. The choice of a combined test may rest on several factors, such as the information available (e.g. only p values may be available in some instances), ease of computation, or the desire for consistency between the combined test selected and the statistic used for the independent tests (e.g. the Winer procedure would be more consistent with summing independent t-statistics,

while the Stouffer procedure would be more consistent with summing independent z-statistics). Measures of effect size are clearly valuable in providing potential insight into the differential impact of a given program, information that generally is more obscured when relying solely upon statistical tests.

References

Adcock, C.J. A note on combining probabilities. Psychometrika, 1960, 25, 303-305.

Birnbaum, A. Combining independent tests of significance. Journal of the American Statistical Association, 1954, 49, 554-574.

Campbell, D.T. Can we be scientific about policy research? Award address presented at the meeting of the American Educational Research Association, New York, March 1982.

Campbell, D.T. & Stanley, J.C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Carkhuff, R.R. Helping and human relations: A primer for lay and professional helpers (Vol. I). New York: Holt, Rinehard & Winston, 1969.

Cohen, J. Statistical power analysis for the behavioral sciences (Rev. ed.), New York: Academic Press, 1977.

Cook, T.D. & Grinder, C. Metaevaluation research. In L. Sechrest et al (eds.). Evaluation Studies Review Annual (Vol. 4). Beverly Hills: Sage, 1979.

CTB/McGraw-Hill. Comprehensive tests of basic skills. Levels 1-3, Forms S & T. Monterey, CA.: 1973/1975.

Engler, C.M., Saltzman, G.A., Walker, M.L. & Wolf, F.M. Medical student acquisition and retention of communication and interviewing skills. Journal of Medical Education, 1981, 56, 572-579.

Fisher, R.A. Statistical methods for research workers (4th ed.). London: Oliver and Boyd, 1932, pp. 99-101.

Fisher, R.A. Combining independent tests of significance. American Statistician, 1948, 2 (5), 30.

Glass, G.V. Primary, secondary, and meta-analysis of research. Educational Research, 1976, 5, 3-8.

Glass, G.V. Integrating findings: The meta-analysis of research. In L.S. Shulman (Ed.) Review of Research in Education (Vol. 5). Itasca, Illinois: F.E. Peacock, 1978.

Glass, G.V., McGaw, B. & Smith, M.L. Meta-Analysis in social research. Beverly Hills, CA.: Sage, 1981.

Kiresuk, T.S. & Lund, S.H. Process and outcome measurement using goal attainment scaling. In Evaluation Studies Review Annual: (Vol. 1), G.V. Glass (ed.). Beverly Hills:Sage, 1976.

Koziol, J.A. & Perlman, M.D. Combining independent Chi-squared tests. Journal of the American Statistical Association, 1978, 73, 753-763.

Littell, R.C. & Folks, J.L. Asymptotic optimality of Fisher's method of combining independent tests II, Journal of the American Statistical Association, 1973, 68, 193-194.

McGaw, B. & Glass, G.V. Choice of the metric for effect size in meta-analysis. American Educational Research Journal, 1980, 17, 325-337.

Mosteller, F.M. & Bush, R.R. Selected quantitative techniques. In Handbook of social psychology: Vol. 1. Theory and method, G. Lindzey (ed.). Cambridge, Mass.: Addison-Wesley, 1954.

Pearson, K. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random, Biometrika, 1933, 25, 379-410.

Rosenthal, R. Combining results of independent studies. Psychological Bulletin, 1978, 85, 185-193.

Rosenthal, R. On telling tails when combining results of independent studies, Psychological Bulletin, 1980, 88, 496-497.

Saltzman, G.A., Wolf, F.M., Savickas, M.L. & Walker, M.L. Dogmatic thinking and communication skills of student physicians. Psychological Reports, 1981, 48, 853-854.

Stouffer, S.A. et. al. The american soldier: Vol. 1. Adjustment during army life. Princeton: Princeton University Press, 1949, p. 45, footnote 15.

Van Zwet, R.R. & Oosterhoff, J. On the combination of independent test statistics. Annals of Mathematical Statistics, 1967, 38, 659-680.

Winer, B.J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971, pp. 49-50.

Wolf, F.M. Summary of interviewing skills for classes of 1981, 1982, 1983. Rootstown: Behavioral Sciences Program, Northeastern Ohio Universities College of Medicine, 1981.

Wolf, F.M. & Blixt, S.L. Evaluation of the Madison Local Schools Title IV-C model math project: Final report. Kent, Ohio: Bureau of Educational Research & Services, Kent State University, 1979.

Wolf, F.M. & Blixt, S.L. A cross-sectional cross-lagged panel analysis of mathematics achievement and attitudes: Implications for the interpretation of the direction of predictive validity. Educational & Psychological Measurement, 1981, 41, 829-834.

Wolf, F.M. & Blixt, S.L. The use of goal attainment scaling in the evaluation of community health services. Mid-Western Educational Researcher, 1981, 2 (1), 35. (Abstract)

Wolf, F.M. & Spies, C.J. Assessing the consistency of cross-lagged panel effects with the Fisher combined test. American Statistical Association 1981 Proceedings of the Social Statistics Section, 1981, 24, 506-511.

Table 1

Means, Standard Deviations, and Paired t-Tests
for Student Performance on the
CTBS Mathematics Achievement Test

Grade	n	Pre		Post		Paired
		M	Sd	M	Sd	t
1	309	0.9	.62	2.4	.51	43.32*
2	362	2.1	.64	3.1	.72	35.47*
3	363	3.1	.75	4.6	1.21	36.11*
4	340	4.4	1.33	5.6	1.52	24.39*
5	331	5.4	1.55	6.7	2.08	19.24*
6	304	6.2	1.93	7.5	2.21	17.30*
7	322	7.2	2.09	8.5	2.49	18.04*
8	299	8.2	2.29	9.4	2.47	14.17*

* $p < .001$, two-tailed test

Table 2

Results of Paired t-Tests for Student Performance
on the CTBS Mathematics Achievement Test

Grade	n	Paired t	One-tailed p	$-2 \log_e p$	d	U_3(%)
1	309	43.32	.0005	15.20	2.42	99.6
2	362	35.47	.0005	15.20	1.56	94.1
3	363	36.11	.0005	15.20	2.00	97.7
4	340	24.39	.0005	15.20	0.90	81.6
5	331	19.24	.0005	15.20	0.84	79.5
6	304	17.30	.0005	15.20	0.67	74.9
7	322	18.04	.0005	15.20	0.62	73.2
8	299	14.17	.0005	15.20	0.52	69.8
Average:					1.19	83.8

Table 3

Means, Standard Deviations, and Paired
t-Tests for First-year Medical Student Performance
on Carkhuff Standard Index of Communication

Graduation Year	n	Pre M	Pre Sd	Post M	Post Sd	Paired t
1981	46	1.55	.30	2.60	.22	-24.18*
1982	44	1.32	.39	2.54	.48	-14.16*
1983	42	1.47	.52	2.55	.59	-8.55*

* $p < .001$, two-tailed test

Table 4

Results of Paired t-Tests for First-year Medical Student Performance on Carkhuff Standard Index of Communication

Graduation Year	n	Paired t	One-tailed p	$-2 \log_e p$	d	U_3(%)
1981	46	-24.18	.0005	15.20	3.52	99.9
1982	44	-14.16	.0005	15.20	3.12	99.9
1983	42	-8.55	.0005	15.20	2.07	98.0
Average					2.90	99.8

Table 5

Goal Attainment Scaling Evaluation Results
for County Mental Health Agency Services

Service	n	Intake M	Intake Sd	Follow-up M	Follow-up Sd	Paired t
Adult Mental Health Services	20	37.62	3.95	58.04	6.83	12.02*
Elderly Home Aid Services	19	34.68	4.82	53.93	8.51	10.28*
Crisis Intervention/Hotline	20	28.49	4.25	45.97	6.61	10.57*
Children's Services	31	38.13	10.52	57.78	9.66	11.15*

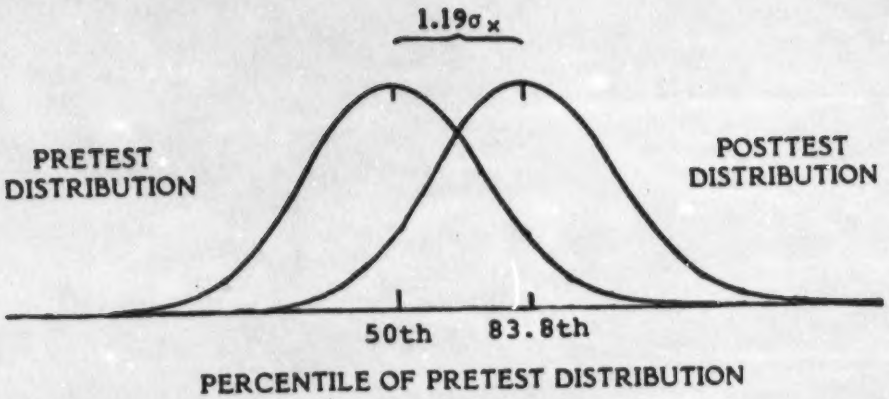
* $p < .001$, two-tailed test

Table 6

Results of Paired t-Tests for Goal Attainment Scaling Evaluations of County Mental Health Agency Services

Service	n	Paired t	One-tailed p	$-2\log_e p$	d	U_3(%)
Adult Mental Health Services	20	12.02	.0005	15.20	5.17	99.9
Elderly Home Aid Services	19	10.28	.0005	15.20	3.99	99.9
Crisis Intervention/Hotline	20	10.57	.0005	15.20	4.11	99.9
Children's Services	31	11.15	.0005	15.20	1.87	96.9
Average:					3.79	99.9

Figure 1. Illustration of average effect size in standard deviation units (σ_x) of student performance on the CTBS Mathematics Achievement Test for grades 1-8.


1.19σx
PRETEST DISTRIBUTION
POSTTEST DISTRIBUTION
50th
83.8th
PERCENTILE OF PRETEST DISTRIBUTION